Exploring Large Language Models for Evaluating Automatically Generated Questions

Jeffrey S. Dittel[1], Michelle W. Clark[1], Rachel Van Campenhout[1], Benny G. Johnson*[1]

[1] *VitalSource Technologies, Raleigh, North Carolina, USA*

Abstract

Automatic question generation has emerged as an effective and efficient method for incorporating formative practice into electronic textbooks on a large scale. This advancement, however, introduces new challenges in ensuring the quality of the generated questions. Traditionally, analyzing student responses has been effective in identifying low-quality questions. However, preemptively filtering out substandard questions before they reach students would be more desirable. In this study, we present preliminary findings on a promising technique that leverages a large language model (LLM) to identify potentially low-quality questions. Our hypothesis is that questions an LLM fails to answer correctly may contain quality issues, particularly since LLMs generally outperform students in answering automatically generated questions. Using a data set of questions from an open-source textbook, our method successfully identified nearly 30% of the questions that were rejected through analysis of student answer data. These results suggest that LLMs can be a valuable tool in improving the quality control process of automatically generated questions.

Keywords

Automatic question generation, artificial intelligence, formative practice, question evaluation, content improvement service, iterative improvement

1. Introduction

Using artificial intelligence for automatic question generation (AQG) has emerged as an effective and efficient way to add formative practice at scale to electronic textbooks. Placing formative practice alongside the electronic text (etext) significantly enhances learning outcomes by implementing a learning science principle known as the doer effect. The doer effect is proven to be beneficial for all students and has approximately six times the impact on learning than reading alone [1, 2, 3]. Since 2022, more than 2.5 million automatically generated (AG) questions have been placed into over 9,000 etexts as a free learning feature, named CoachMe, within VitalSource's Bookshelf electronic reader platform.

This practice feature contains several types of AG questions, including fill-in-the-blank (FITB), matching, multiple choice, and free response. The FITB questions, which comprise the majority of the AG questions, are the focus of the present study. For details of the AQG process

* Corresponding author.
jeff.dittel@vitalsource.com (J. S. Dittel); michelle.clark@vitalsource.com (M. W. Clark); rachel@acrobatiq.com (R. Van Campenhout); benny.johnson@vitalsource.com (B. G. Johnson)
0000-0002-4913-4427 (J. S. Dittel); 0009-0002-1500-9166 (M. W. Clark); 0000-0001-8404-6513 (R. Van Campenhout); 0000-0003-4267-9608 (B. G. Johnson)

used, see [4]. As formative practice, students are allowed as many attempts to answer as they like, receiving immediate feedback, and can also reveal the answer if stuck.

Research on questions generated through this artificial intelligence (AI) process found that they performed equally as well with students as human-authored questions on several key metrics [5, 6]. However, as anyone who has created educational content is aware, no content is perfect, and it is inevitable that problems or errors occur. Just as no human could write millions of perfectly performing questions, neither does AI always generate perfect questions. Using AQG on this unprecedented scale presents a new challenge in how to monitor and perform quality assurance on this enormous question set.

The Content Improvement Service (CIS) is an automated adaptive system that was developed in response to this practical need [7, 8]. The CIS is a platform-level system that monitors real-time clickstream data for all questions delivered in all e-textbooks. The CIS evaluates question quality using a variety of modular plug-in tools and operates independently without requiring human involvement. There are currently two primary evaluation methods used. One is a Bayesian analysis that removes questions with a mean score likely to be below a minimum acceptable threshold. The other uses student feedback given through a thumbs up or down rating mechanism. Questions with more than one thumbs down rating within the first 100 students answering are removed. For more details on the CIS analysis methods, see [8].

Even though the CIS can detect unacceptable questions using a relatively small amount of data [8], it would be more desirable to preemptively remove these before releasing them to students. To this end, post-AQG human assessment of question quality is common due to the well-known limitations of AG questions. From a review of the AQG evaluation literature by Kurdi et al. [9], studies that reported on "question acceptability or overall quality" from human review found percentages of acceptable questions ranging from 50% to 93%. Early in the CoachMe release, a human review pass was performed by the AQG development team to check for common AQG quality issues that were not subject matter-related and did not require pedagogical expertise. For details of this review process, see [4].

There were practical difficulties with the human review process, however. For one, it was not scalable to keep up with the large volume of questions needing review. For expedience, it was also necessary to review the questions in isolation in spreadsheets outside of the etext, i.e., without the context of the textbook material for which the questions serve as practice. The pace at which questions could be reviewed would have been much slower within the etext. While isolated review increased the volume of questions that could be reviewed, this potentially negatively affected review quality. Specifically, this led to retaining some questions that may not have been considered acceptable with the benefit of contextual information. Furthermore, reviewers were not subject matter experts (SMEs) in most domains they reviewed. It was not practical to recruit SMEs (e.g., higher education instructors) for question review across multiple disciplines at the scale needed. In a regression model on a large question usage data set from a recent study [10], whether a question had passed a manual review (under the conditions described) did not have a statistically significant impact on whether a student rated the question as not helpful.

Not surprisingly, the manual review step was discontinued as the scale of CoachMe increased. However, the major difficulties with manual review, namely scale, inability to consider context, and lack of reviewer subject matter expertise, could potentially be mitigated or eliminated by incorporating a large language model (LLM) into an automated review process.

Volume is not an issue for an LLM, and both time and cost are significantly reduced compared to human review. Unlike with human review, it is practical (and prudent) to give an LLM contextual information to consider. And, given their massive training data sets, it seems reasonable that an LLM could have greater subject matter expertise in many domains than a non-SME human reviewer. For example, as seen in the Results and Discussion section, the LLM used in this work was able to answer CoachMe questions correctly at a much higher rate than students.

Use of an LLM may therefore have potential to meaningfully enhance automated question review and reduce the number of unacceptable questions that must be detected through collection and analysis of student data by the CIS. In this work, we take a first step in investigating this possibility. Our hypothesis is questions that an LLM fails to answer correctly may contain issues that would lead to their rejection based on analysis of student answer data. To be clear, here we are using an LLM to attempt to improve assessment of the quality of questions generated by the CoachMe AQG process; CoachMe does not yet use LLMs in the question generation process itself.

2. Methods

The data set in this study is from use of CoachMe in the Chemistry 101 course at a U.S. major public university in the Fall 2023 semester. A total of 744 students were enrolled in the course. The textbook used in the course was *Chemistry: Atoms First* [11] from OpenStax. The textbook contained 282 AG formative practice questions added by CoachMe, of which 106 were FITB cloze questions. The implementation of CoachMe by the course instructors was that students who answered 75% of the questions in relevant chapters (regardless of correctness) would have an additional reading quiz dropped at the end of the semester.

Since the CIS rejects questions receiving more than one thumbs down within the first 100 students, questions having at or near 100 answering students were selected for analysis, using a natural break in the data observed at 88 students. (The mean score analysis can typically reach a decision on question rejection in fewer than 100 students.) This resulted in 54 of the 106 FITB questions being selected. The mean number of students answering each question was 118.3 and the overall question mean score on the first attempt was 41.2%. A total of 247 students answered questions in the data set. The mean number of questions answered per student was 25.9, with 71 students answering all 54 questions in the data set. Of these 54 questions, 27 (50%) would be flagged for rejection using one or both CIS criteria (21 by mean score and 13 by student ratings). Note that all selected questions had passed the human review process, even though half ended up being rejected by the CIS analyses.

The LLM evaluation work was done using GPT-4 [12]. The first evaluation performed was simply to direct the LLM to answer the question as if it were a student. GPT-4's temperature parameter was set to 0 (the lowest value), which causes it to respond with the answer word that is most probable under the model. GPT-4 uses a system prompt to provide instructions for the conversation, such as setting a role for the LLM to assume, and a user prompt that provides the actual query. The system prompt used was "`You are a college student with a 4.0 GPA.`" The user prompt, illustrated using one of the questions from the data set, was

```
Here is a fill-in-the-blank question from your textbook. Please answer with
the word that best fits in the blank. Answer only with a single word.
```

```
Question: The uncertainty principle can be shown to be a consequence of
wave-particle duality, which lies at the heart of what distinguishes modern
______ theory from classical mechanics.
Answer:
```

For the above question, GPT-4 responded with “quantum”, which is correct. Answering a question incorrectly was taken as a predictor that the question will be rejected by the CIS. Why might this simple criterion be useful? Since the LLM is much better at answering the questions than students, an incorrect answer may be due to a defect in the question rather than a limitation in the LLM’s knowledge; examples are presented below. Precision and recall were calculated for this criterion.

When CoachMe is used as intended, students answer the questions while reading the textbook content; this leads to improved learning via the doer effect [1, 2, 3]. The evaluation was therefore repeated providing the LLM with the complete paragraph in which the question’s sentence appeared, with the same answer word removed. This is intended to resemble how students should use the formative practice more closely, i.e., answering the questions immediately after reading the relevant textbook content.

3. Results and Discussion

When the LLM was given each question to answer without additional context, 43 of 54 (79.6%) were answered correctly. This is in stark contrast with the first attempts by students, which were only 41.2% correct. A z test of two proportions shows this difference is statistically significant (p << .001).

Taking an incorrect answer as predicting rejection by the CIS had precision 72.7% and recall 29.6%. This precision is perfectly acceptable. Maximizing precision is not a pressing concern since the AQG process generates more questions than needed, with the rest held in reserve as replacements for questions rejected by the CIS [8]. Recall is also reasonable considering the minimal effort involved in obtaining it, i.e., merely checking if the LLM answers the question incorrectly (more on this topic below).

An example of a question correctly predicted for rejection by the CIS is

```
The order of a(n) ______ bond is a guide to its strength; a bond between
two given atoms becomes stronger as the bond order increases.
```

The correct answer is “covalent” and the LLM’s answer was “chemical”. While both are reasonable words for completing the sentence in isolation, “covalent” is more specific to the textbook context on the topic of nolecular orbitals for diatomic molecules, which concerns covalent bonding.

Another question correctly predicted for rejection is

```
One particularly characteristic ______ of waves results when two or more
waves come into contact: They interfere with each other.
```

The correct answer word (i.e., appearing in the textbook sentence) is “phenomenon” while the LLM answered “property”. Here, these words are synonymous, completing the sentence equally well (even considering context), but “property” would be counted as incorrect. This illustrates how the LLM-based answer criterion can be useful, by identifying when an equally good answer word as the one used by the textbook author exists.

Providing the textbook paragraph in which the sentence occurred as context is expected to increase the proportion of questions correctly answered. This was observed, with 46 of 54 (85.2%) correct, compared to 79.6% correct without this contextual information. More correctly answered questions means fewer questions predicted as rejected, and so recall should decrease and precision increase. While this was the case, with precision 75.0% and recall 22.2%, interestingly, the difference made by this additional information was small.

Although many questions rejected by the CIS were not identified by this method, it is important to note that it is only one of several used to identify unacceptable questions without student data. The focus is detecting cases where the answer word is not sufficiently predictable from the question and the LLM’s background knowledge of the subject. This is only one way AG questions can prove unacceptable [4, 9], and thus high recall of rejected questions is not necessary as a measure of success. An example question that the LLM answered correctly but was given thumbs down by multiple students illustrates this point.

```
Because a hydrogen ______ molecule contains two oxygen atoms, as opposed
to the water molecule, which has only one, the two substances exhibit very
different properties.
```

The LLM correctly answered “peroxide”, which is highly predictable in this context, and thus the question was not predicted for rejection. However, students viewed the question as not helpful because this sentence was serving as an example to illustrate a central concept (the chemical mole concept), and not as an important fact that needed to be retained. This reason was not related to the answer word’s predictability.

We have investigated a very simple but powerful method of identifying unacceptable AG questions without student data. Preliminary results are promising, identifying almost 30% of questions that would be rejected by data analysis if released to students. Directions for continued work involve analyzing a larger and more varied question sample, improving precision and recall by involving the LLM more directly in the question generation process (e.g., in selecting the answer words, not just assessing them post-AQG), and extension to other AG question types like multiple choice.

The data for this study (AG questions, anonymized student interaction events, and LLM analysis results) are available at our open-source data repository [13].

References

[1] K. R. Koedinger, J. Kim, J. Z. Jia, E. A. McLaughlin, and N. L. Bier, Learning is not a spectator sport: Doing is better than watching for learning from a MOOC, in Proceedings of the Second ACM Conference on Learning @ Scale (L@S '15), 2015, pp. 111–120. doi: 10.1145/2724660.2724681.

[2] K. R. Koedinger, E. A. McLaughlin, J. Z. Jia, and N. L. Bier, Is the doer effect a causal relationship? How can we tell and why it's important, in Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16), Edinburgh, United Kingdom, 2016, pp. 388–397. doi: 10.1145/2883851.2883957.

[3] R. Van Campenhout, B. Jerome, and B. G. Johnson, The doer effect at scale: Investigating correlation and causation across seven courses, in LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), 2023, pp. 357–365. doi: 10.1145/3576050.3576103.

[4] B. G. Johnson, R. Van Campenhout, B. Jerome, M. F. Castro, R. Bistolfi, and J. S. Dittel, Automatic question generation for Spanish textbooks: Evaluating Spanish questions generated with the parallel construction method, International Journal of Artificial Intelligence in Education, vol. 34, no. 2, pp. 123-137, Apr. 2024. doi: 10.1007/s40593-024-00394-1.

[5] R. Van Campenhout, J. S. Dittel, B. Jerome, and B. G. Johnson, Transforming textbooks into learning by doing environments: An evaluation of textbook-based automatic question generation, in Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education CEUR Workshop Proceedings, 2021, pp. 1–12. [Online]. Available: https://ceur-ws.org/Vol-2895/paper06.pdf.

[6] B. G. Johnson, J. S. Dittel, R. Van Campenhout, and B. Jerome, Discrimination of automatically generated questions used as formative practice, in Proceedings of the Ninth ACM Conference on Learning@Scale, 2022, pp. 325–329. doi: 10.1145/3491140.3528323.

[7] B. Jerome, R. Van Campenhout, J. S. Dittel, R. Benton, S. Greenberg, and B. G. Johnson, The Content Improvement Service: An adaptive system for continuous improvement at scale, in Interaction in New Media, Learning and Games. HCII 2022. Lecture Notes in Computer Science, A. Meiselwitz et al., Eds. Cham: Springer, 2022, vol. 13517, pp. 286–296. doi: 10.1007/978-3-031-22131-6_22.

[8] B. Jerome, R. Van Campenhout, J. S. Dittel, R. Benton, and B. G. Johnson, Iterative improvement of automatically generated practice with the Content Improvement Service, in Adaptive Instructional Systems. HCII 2023. Lecture Notes in Computer Science, R. Sottilare and J. Schwarz, Eds. Cham: Springer, 2023, pp. 312-324. doi: 10.1007/978-3-031-34735-1_22.

[9] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, A systematic review of automatic question generation for educational purposes, International Journal of Artificial Intelligence in Education, vol. 30, no. 1, pp. 121–204, 2020. doi: 10.1007/s40593-019-00186-y.

[10] B. G. Johnson, J. S. Dittel, and R. Van Campenhout, Investigating student ratings with features of automatically generated questions: A large-scale analysis using data from natural learning contexts, accepted for publication in EDM 2024: The 17th International Conference on Educational Data Mining, Atlanta, Georgia, USA, July 14–17, 2024.

[11] P. Flowers, E. J. Neth, W. R. Robinson, K. Theopold, and R. Langley, Chemistry: Atoms First, 2nd ed. Houston: OpenStax, 2019.

[12] OpenAI, GPT-4 Technical Report, ArXiv, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774.

[13] VitalSource Technologies, VitalSource Supplemental Data Repository, 2024. [Online]. Available: https://github.com/vitalsource/data.